从原材料到资产
——数据资产化的挑战和思考

吴 超

中国科学院计算机网络信息中心 北京 100190

摘要 嵌入式和可穿戴设备正普及大众，各类传感器已可对用户敏感数据采集，无处不在的互联网和普及的云计算以及存储设施，也使得传输和管理这些数据变得越来越容易，深度学习等模型也开始充分挖掘这些数据的价值；然而数据从一开始作为原材料，到最后成为产品提供给用户，其中需要经历一系列的加工和增值过程，在此过程中经济因素将成为最大的推动力量。文章讨论了数据资本化的问题，在此过程中要推动从数据到数据产品的价值链，很多关键的经济问题需要考虑，其中核心问题包括数据作为资产的定价问题，以及隐私保护等。

关键词 数据资产化，数据定价，隐私保护

DOI 10.16418/j.issn.1000-3045.2018.08.004

计算技术和能力已经完全普适化，对数据的观察和整合、分析和解释，正在不断创造新的知识，推动着科学技术的进步和社会的发展。嵌入式和可穿戴设备正普及大众，各类便携传感器已可对用户敏感数据进行采集，如智能手机包含了GPS、加速度计、距离及光线传感器、摄像头、陀螺仪、指纹传感器，甚至还包含心率监测器等数据采集和感知设备。无处不在的互联网和普及的云计算、存储设施，也使得传输和管理这些所采集数据变得越来越容易。对这些所采集数据可从两个方面进行利用：① 建立数据的统计模型以帮助公共和私人部门了解社会运行各方面的整体情况，如流行病的早期检测；② 从微观层面提供个性化服务，如对每个居民提供产品和服务推荐。

在深度网络出现之前，机器学习模型无需大量训练数据，就算有更多数据，模型也不能训练得更好（模型进入 saturation 状态）[1]；而对深度网络来说，因为其足够深，需要训练的参数足够多，所以它对数据是饥饿的——当数据越多的时候，能构建的网络就越深，其性能就越好，这是大数据的作用。如今，这种以大数据+深度神经网络为代表的人工智能技术，正在深远地影响着社会生活的各个方面。而数据作为一种原材料，通过数据分析建模的加工挖掘，能产生新的价值，已成为新的生产力来源和资产。

众多案例已展示了数据的应用价值[2,3]，然而一个技术要深刻地推进社会发展，它需要从具有应用价值发展

修改稿收到日期：2018年8月17日

为具有应用+经济的双重价值。从经济价值的眼光来看大数据，我们可以看到所谓的“数据”在整条价值链上处在起点的位置。数据从一开始作为原材料，到最后成为产品提供给用户，其中经历了一系列的加工和增值过程，包括清理[4]、语义化[5]、融合[6]、分析[7]、建模[8]、知识提取[9]、应用[10]、分发[11]等关键步骤，如同一个工业产品，从原材料到最终产品形态再到市场，是一个复杂的价值链，需要精巧的协同工作。而在目前大部分的大数据研究中，关注点还仅停留于这些具体过程的技术基础，我们相信随着整个生态环境的建立，每个步骤背后的经济因素将成为最大的推动力量。

1 数据资产化中的隐私保护

在数据资产化过程中，隐私保护成为关键问题。数据所有权和隐私权问题长期以来都是信息产业的核心问题[12]。隐私可视为用户对信息流通程度和方式的控制权。传统隐私保护研究较关注访问控制及数据发布前去除个人信息，并防止多个数据源融合之后恢复所去除的个人信息。而随着大数据、移动采集设备和机器学习等技术发展，在数据收集阶段进行隐私保护，是面临的一个新问题。

由于数据对于构建高效模型越来越重要，数据收集中的隐私保护应处在一种权衡取舍状态。解决隐私保护问题，并不能将其孤立地看待，而是应该放在一个更大的框架中，即在用户的隐私权利和从用户数据中获得服务与资源之间进行权衡取舍，使之在当前情境达到最优。因此，需要建立一个能支持多方双赢的隐私保护机制：一方面保障用户隐私可控而促进数据交易和流通；另一方面促进数据驱动商业模式和生态健康发展。

数据收集作为开发创新及个性化、情境化应用的关键环节，从隐私角度来看，处在“法律灰色地带”。当前，大部分应用程序只标明了其市场价格，而对收集数据的范围和粒度并没有明确的协议。例如，一个导航软件应用系统可在用户不知情的情况下，在后台持续大量收集该用户数据。以移动应用为例，91% 的 IOS 应用程序和 83% 的 Android 应用程序存在至少一种泄露用户隐私的风险行为[13]。Facebook、Apple、Twitter、Yelp、Path 等公司都曾因被指控发布侵犯隐私的移动应用程序而成为诉讼的焦点[14]。

应用程序（特别是移动应用）往往将数据收集信息（如类型、数量）描述的暧昧不明，虽然数据收集通常会在最终用户协议中被提及（如在 Apple App Store 中），但用户通常并不会阅读这些冗长文档，而直接选择同意该条款。况且最终用户协议中的许可声明往往语焉不详，且具误导性，实际中却大量收集用户敏感数据。而且数据收集的隐私保护并不是一个有或无的问题[15]，而是一个程度问题。尽管部分应用程序商店（如 Google Play Store）对应用程序访问用户数据提供了一定的控制机制，但对数据访问的粒度仍然缺乏支持，在 Google Play Store 中标明了应用需要访问的数据类型，对数据收集的数量和频率并不明确，而数据的数量常常是很关键的[16]。

隐私保护与数据效用之间需要妥协和平衡[17-19]，也要在技术方案上构建一种生态环境，在这种情况下，各国政府出台了一系列政策法规。例如，欧洲的数据保护政策 General Data Protection Regulation（GDPR），已于 2018 年 5 月开始实施。Determann[20]讨论了 GDPR 与其他国家隐私保护规范的差异。Post[21]分析了 Google 在欧盟（西班牙）收到隐私侵犯调查及此事件带来的深远影响，以及引起欧盟后续的法律环境变化。2017 年 6 月 1 日正式实施的《中华人民共和国网络安全法》，强调了中国境内网络运营者对所收集到的个人信息所应承担的保护责任和违规处罚措施。但专项个人信息保护法现尚在制订中。

2 数据资产化中的数据定价与交易

要推动从数据到数据产品的价值链，还有很多关键的经济问题需要考虑，其中一个核心的问题是数据作为

资产的定价问题。数据与其他原材料在4个方面有很大不同：① 数据的使用不会带来数据的消耗，数据的开发不是排他的，甚至反而是利他的；② 聚合后的数据比单独的数据更有价值，也应该具有更高的价格；③ 同样种类的数据，不同来源的数据具有不同的价值，这点在医疗数据中尤为突出；④ 同样的数据在不同的使用者看来，也是价值各异。在这些特殊的条件，如何对数据资产进行定价是一个很难的问题，我们认为采用一种基于市场协商的价格或许更为现实可行。

目前大部分应用程序正在从以广告收入为主的商业模式向基于个人数据采集的商业模式过渡。但在当前的数据收集模式下，用户无法凭借其贡献的数据而获取奖励，这种模式表面上可使应用程序服务从中受益，然而考虑到潜在的法律后果，实际上是阻碍了其商业模式的可持续发展。由于用户数据的所有权不明，导致数据难以有效流通。

非法的数据交易会对个人数据等高价值信息的安全造成影响[22]，对非法数据交易的购买方和协助方都应进行处罚。特别对于定价来说，传统的效用价格论、成本价格论等定价模式并不适用[23]。金融资产的定价理论有值得借鉴的地方，然而供应方提供的数据很难与数据需求方的应用方向精准匹配，供需错配的问题无法解决。另外，需求方在不确定某数据资源是否能真正能给组织带来收益情况下，很难给出一个较高的价格。刘洪玉等[24]认为在大数据交易过程中，由于缺乏足够的历史参考，其数据资源的交易价格很难确定，因此提出一种基于竞标机制的鲁宾斯坦模型，用于大数据交易双方进行讨价还价，以求达成一个交易的均衡价格。Li 和 Miklau[25]提出了数据市场定价的3个原则和定价函数的基本结构。Valz[26]通过数据内容动态调整定价；翟丽丽等[27]从资产的期权价值角度来评估大数据资源的价值，并指出数据在不断变化和更新，加上数据的非独占性等情况的出现，数据资产的价值可能会下降，最后综合这些因素构建了一个评估模型来计算数据资产的价值。市场有助于数据合理定价[28]，Iyilade 和 Vassileva[29]提出了一种隐私保护的数据交易算法，其基本思路是应用程序之间通过市场机制来优化数据共享。

但是，这些定价方式都存在一个共同的问题：对数据交易中的安全问题和隐私泄露等有较大的担忧，大量数据源未被激活[30]。虽然数据具有明显的商品特征，它却有很强的非传统商品属性，如复制成本接近于0、非排他性、时效性等。这造成了近年来，虽然建立了一些数据交易所（如2017年关闭的微软Azure DataMarket），但数据交易仍难以成规模，数据还很难流通并发挥价值。

有了定价，还需要交易。数据资产要产生价值，需要进行流通。早期数据流通研究是从数据可达性、分布式系统可靠性等角度出发的[31]。然而，在数据收集和交易过程中始终存在着“信息不对称”：目前用户缺乏对数据收集的认知，因而始终处于弱势。虽然目前有一些研究提出基于法律和交易的体系解决方法，但缺乏实在的技术方案。我们在 Imperial Festival 和英国数字经济会议上的公众调查所了解到，大多数用户并不清楚自己究竟有多少数据被应用程序收集。

我们提出了一种新的移动隐私保护模型——PBD 模型[32]（Pay-by-Data），PBD 将数据显式地作为一种应用效能的支付手段，用户和数据收集者之间达成收集和反馈的协议，通过保护隐私达到数据的合理定价。

（1）在数据消费者与数据提供者之间引入数据付费协议（data pricing agreement，DPA）。DPA 以数据（隐私）作为计价工具，定义一种新型的应用服务付费方式，允许用户交易自己的数据（隐私）以获取服务或是其他激励。DPA 详细描述应用所访问的数据类型、收集数据的频率以及用户所获得的回报；并针对不同的数据质量，制定不同的价格机制。因此微观用户数据的收集是被数据付费协议显式规范的，减少了肆意侵犯用户隐私的行为。

（2）通过定制的Android等平台，改进应用程序与底层移动服务之间的通信及请求获取用户数据的方式。用

户数据的访问由数据付费认证服务控制，提供了更细的粒度支持。数据付费协议在基于区块链的智能合约上实现，从而保证公平执行和可追溯性。同时提供新的数据访问开发API供应用开发使用。

（3）研究通过市场的机制寻找隐私保护和数据收集之间的平衡。透明可信的数据收集明确定义用户的数据收集所对应的报酬（即资源和服务），产生激励；并因此构建一种数据定价和交易方法，数据被用作一种货币，用来购买应用提供的服务和资源（这里也包括现实货币），通过有效的市场机制，使这些应用程序和用户之间达到定价均衡。

参考文献

1 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research, 2010, (9): 249-256.

2 Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. America: McKinsey Global Institute, 2011.

3 McAfee A, Brynjolfsson E. Big data: The management revolution. Harvard Business Review, 2012, 90(10): 60-66.

4 Rahm E, Hong H D. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000, 23(23): 3-13.

5 Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web. Heidelberg: Springer Berlin Heidelberg, 2007: 11-15.

6 Hall D L, Llinas J. An introduction to multisensor data fusion. Proceedings of the IEEE, 2002, 85(1): 6-23.

7 Trnka A. Big data analysis. European Journal of Science and Theology, 2014, 10(1): 143-148.

8 Wu X, Zhu X, Wu G Q, et al. Data mining with big data. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(1): 97-107.

9 Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: from big data to big impact. Mis Quarterly, 2012, 36(4): 1165-1188.

10 Murdoch T B, Detsky A S. The inevitable application of big data to health care. JAMA, 2013, 309(13): 1351-1352.

11 Viktor M S, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. London: John Murray, 2013.

12 Petrie C. The Proper Use of the Internet: Digital Private Property. IEEE Internet Computing, 2016, 20(2): 92-94.

13 O'Brien K J. Data-gathering via apps presents a gray legal area. New York Times, [2012-10-29]. https://www.nytimes.com/2012/10/29/technology/mobile-apps-have-a-ravenous-ability-to-collect-personal-data.html.

14 van Grove J. Your address book is mine: Many iPhone apps take your data. VB Mobile, [2012-02-14]. https://venturebeat.com/2012/02/14/iphone-address-book/.

15 Xu L, Jiang C, Wang J. Information security in big data: privacy and data mining. IEEE Access, 2014, 2: 1149-1176.

16 Montjoye Y A D, Hidalgo C A, Verleysen M, et al. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, 2013, 3(6): 1376.

17 Xu J, Wang W, Pei J, et al. Utility-based anonymization for privacy preservation with less information loss. Acm Sigkdd Explorations Newsletter, 2006, 8(2): 21-30.

18 Gionis A, Tassa T. k-Anonymization with Minimal Loss of Information. IEEE Transactions on Knowledge & Data Engineering, 2008, 21(2): 206-219.

19 Xu L, Jiang C, Chen Y, et al. Privacy or Utility in Data Collection? A Contract Theoretic Approach. IEEE Journal of Selected Topics in Signal Processing, 2015, 9(7): 1256-1269.

20 Determann L. Adequacy of Data Protection in the EU - General Data Protection Regulation as Global Benchmark for Privacy Laws?[2017-06-23]. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2902228.

21 Post R. Data privacy and dignitary privacy: Google Spain, the right to be forgotten, and the construction of the public sphere. Duke Law Journal, 2018, 67(5): 981-1072.

22 史宇航. 个人数据交易的法律规制. 情报理论与实践, 2016, 39(5): 34-39.

23 刘朝阳. 大数据定价问题分析. 图书情报知识, 2016, (1): 57-64.

24 刘洪玉, 张晓玉, 侯锡林. 基于讨价还价博弈模型的大数据交易价格研究. 中国冶金教育, 2015, (6): 86-91.

25 Li C, Miklau G. Pricing aggregate queries in a data marketplace. Webdb, 2012., 2012.

26 Valz D R. Dynamic pricing models for digital content: US, 9076148, 2015.

27 翟丽丽, 王佳妮, 何晓燕. 移动云计算联盟企业数据资产评估方法研究. 价格理论与实践, 2016, (2):153-156.

28 Fricker S A, Maksimov Y V. Pricing of Data Products in Data Marketplaces. Software Business, 2017, 304: 49-66.

29 Iyilade J, Vassileva J. A framework for privacy-aware user data trading. User Modeling, Adaptation, and Personalization, 2013, 7899: 310-317.

30 杨琪, 龚南宁. 我国大数据交易的主要问题及建议. 大数据, 2015, 1(2): 38-48.

31 Cooper B F, Garcia-Molina H. Peer-to-peer data trading to preserve information. Acm Transactions on Information Systems, 2000, 20(2): 133-170.

32 Wu C, Guo Y. Enhanced user data privacy with pay-by-data model// Santa Clara: IEEE International Conference on Big Data. IEEE, 2013: 53-57.

From Raw Materials to Assets—Challenges and Considerations on Data Capitalization

WU Chao

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract Embedded and wearable devices are becoming pervasive, with various sensors collecting user data. With the infrastructure of Internet and cloud computing, it is now much easir to transfer and manage these data. And with deep learning, we can fully mine the value in data. Nevertheless, data needs to be processed with a long workflow, from raw material to final product. Within this workflow, the economic factor would be the most significant force. Therefore, in this article, we discuss the issues in data capitalization. To move data from raw material to final product, we need to consider many aspects, including its pricing, and privacy protection.

Keywords data capitalization, data pricing, privacy protection



吴 超 中国科学院计算机网络信息中心正高级工程师，伦敦帝国理工学院研究员。主要研究方向为面向智慧城市和医疗的数据分析建模方法。E-mail: chao.wu@imperial.ac.uk

WU Chao Senior engineer in Computer Network Information Center, Chinese Academy of Sciences (CAS). Research fellow in Imperial College London. His main research area is about data analysis and modelling for smart city and health care. E-mail: chao.wu@imperial.ac.uk

■责任编辑：张帆